基于地理坐标的微博事件检测与分析*

李进华　安仲杰
(华中师范大学信息管理学院　武汉 430079)

摘要:【目的】利用数据挖掘算法,从海量繁杂的微博数据中检测出有价值的事件信息。【方法】针对国内具有代表性的微博网站,通过使用微博网络开放接口高效收集带有地理坐标的微博数据。使用 K-means、KNN 和决策树三种数据挖掘算法,根据微博数据的发布数、转发数、评论数、用户活跃度和移动强度 5 个指标构建微博的地理规律性特征。将日常地区性的微博数据特征与该地区微博特征的地理规律性进行比较,从而检测出该区域是否有事件发生。【结果】以 2015 年 4 月 15 日、16 日的微博数据作为测试语料,使用文中提出的微博事件检测框架,成功检测到"北京沙尘暴"事件。【局限】在抽取微博地理规律性特征时,采用的样本数据偏少,一定程度上影响了事件检测框架的效果。【结论】基于地理坐标的微博事件检测框架是切实有效的,分析出的事件信息不仅可以帮助用户获取感兴趣的事件资讯,而且可以协助政府部门进行舆情管控和行政决策。

关键词: 微博　事件检测　可视化分析　地理坐标分析
分类号: G354

1　引　言

在当今的信息时代,互联网用户兼有信息接收者和创造者的双重角色。互联网的发展促使网络服务社会化,即网络服务从单一化走向多元化[1]。社交网络代表各种社会关系,它把素未谋面的陌生人、具有血缘关系的亲人以及具有工作关系的同事等组织在一起。通过社交网络平台,用户可以相互交流沟通,进而具有共同价值观、兴趣爱好、理想信念的人形成了许多人际关系圈。随着社交网络的发展,微博应运而生,开创了社交网络服务的新纪元。用户可以通过微博平台构建的单向、双向关注关系进行信息的传播、获取和分享。微博用户可以通过手机、电脑等客户端,以移动 WAP 服务、网页浏览器、即时通讯 IM 软件、手机短信等方式,发布 140 字以内(包含标点符号)的文本、图片及视频等信息,从而实现信息的即时共享。据统计,用户发布的博文形式多种多样,有 69.0%的博文带有图片内容,8.6%的博文含有短链接,还有部分用户进行视频分享,这些多媒体信息丰富和延伸了微博。微博是一个实时的广播平台,用户可以及时接收来自被关注人的微博信息,若用户对某条微博信息感兴趣,可以对该微博进行评论和转发,用户本身也可以对自己的听众广播微博信息。微博被不断转发,尤其是经过意见领袖[2]转发后,信息的传播范围会呈几何级数扩大,产生"裂变式"的信息传播效应。

本文研究的微博事件是指被微博用户发布到微博平台,并且引发了大量用户的转发、评论,产生了较大的社会影响力的事件。近年来,关于微博的研究越来越多,但关于微博事件的研究还相对较少。国内外微博事件检测的研究主要集中在微博事件情感分析[3-4]、微博事件传播和舆情控制研究[5-6]、微博事件检测与追踪[7-8]、微博事件意见领袖识别[9-10] 4 个方面。其中微博事件检测和追踪的研究集中在微博

通讯作者: 李进华, ORCID: 0000-0002-5381-9824, E-mail: lijh@mail.ccnu.edu.cn。

*本文系国家社会科学基金项目"语义网络环境下面向数字化科研的分布式知识发现研究"(项目编号:11BTQ040)和华中师范大学中央高校基本科研业务费专项资金项目"基于统计本体学习方法的文本领域本体自动抽取与演化研究"(项目编号:CCNU13A05048)的研究成果之一。

文本特征选取[11]、微博信息相似度计算算法优化[12]、话题聚类算法的改进[13]、微博事件摘要抽取技术[14]等方面，基本原理是对微博内容进行分词，特征提取、聚类分析，从而挖掘出热点话题。不同于传统的网页和博客，由于微博内容短小的特征，很难提取出足够的信息来判断是否有突发事件发生，因此基于微博内容的突发事件检测技术存在一定的局限性。本文通过主流统计分析软件 R①，从微博的地理分布数据与特征出发对微博热点事件进行跟踪、获取以及可视化分析。

2 微博事件地理信息获取

为了获取到微博用户发布的微博内容，可以通过调用新浪微博的 place/nearby_timeline 接口，获取到某个位置周边的动态微博信息。place/nearby_timeline 接口返回的最大搜索半径是 11 132 米(约为 11 公里)，显然这样能收集到的微博数据是十分有限的。为了解决这个问题，可以通过不断设定不同的经度(long 参数)和纬度(lat 参数)坐标，收集每个经纬度坐标点附近的数据，这样完全可以采集到足够的微博内容，如图 1 所示:

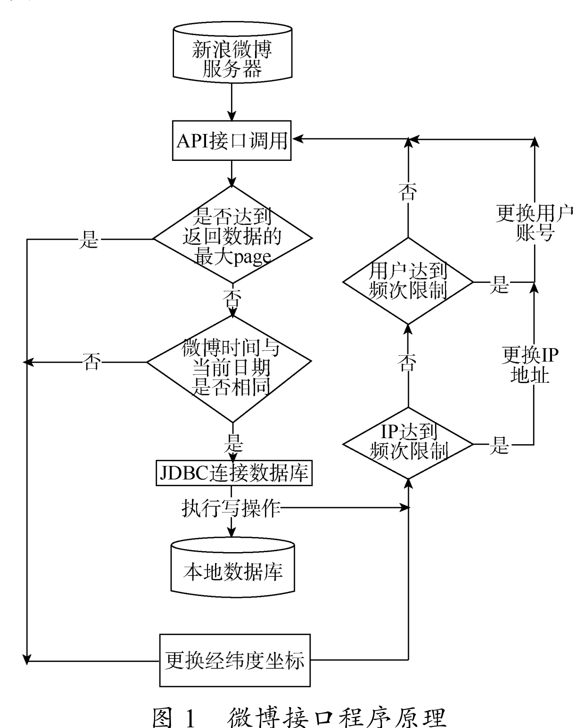


图 1 微博接口程序原理

本文用搜索圆的内接正方形对微博内容待收集区域进行划分。根据圆内接正方形的性质，若外接圆的半径为 11 公里的话，那么内接正方形的边长约为 16 公里。

图 2 以北京市为例，说明地理微博数据采集原理。北京位于东经 115.7°-117.4°，北纬 39.4°-41.6°，东西宽约 160 公里，南北长约 176 公里，因此纬度方向上大约需要 10 次坐标设定，经度方向上需要约 11 次坐标设定，要覆盖整个北京市约需要 110 次坐标设定，部分数据如表 1 所示。

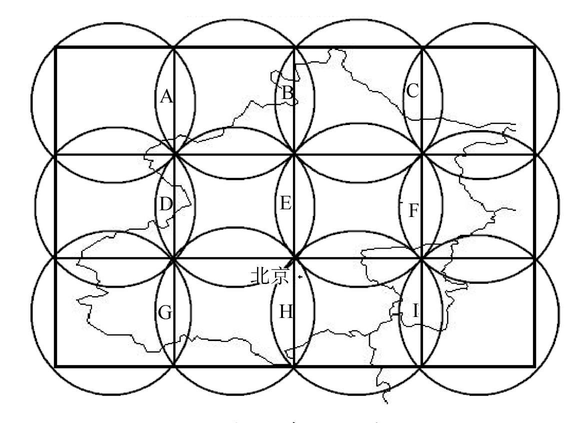


图 2 新浪微博数据采集原理

表 1 北京市部分微博数据收集坐标中心

经度:维度	经度:维度
115.719336:41.458578	117.030503:41.034312
115.718926:41.317156	116.650735:40.751468
116.843019:41.034312	116.096746:41.458578
116.464051:40.751468	116.095516:41.317156
115.908041:41.458578	117.217987:41.034312
115.907221:41.317156	116.837419:40.751468

根据上文采用的微博数据采集方案，同一条微博信息可能被重复收集，存在数据冗余的问题(如图 2 的 ABCDEFGHI 区域的微博)，占用了大量不必要的磁盘空间，其次重复的数据还会对系统的性能和事件检测效果的有效性带来很大影响。在微博数据属性中，微博 ID 是可以唯一标识一条微博的字段，可以通过保持该字段的唯一性(Unique)，筛选掉重复的微博信息，最终得到实验所需数据，部分数据如表 2 所示。

①https://www.r-project.org/.

表 2　地理微博数据

用户 ID	发布时间	纬度	经度	转发数	评论数	粉丝数	微博 ID	微博内容
2142578445	2015-04-16 12:56:56	39.63398	116.32463	0	0	81	3832238061664418	听说今天秀一下北京的蓝天会有好多人点赞? http://t.cn/z8AUYaH
5578044924	2015-04-16 08:59:38	39.4442	116.3018	0	0	16	3832178343107815	早安[太阳] http://t.cn/RACLQ8g
2639854301	2015-04-16 08:47:29	39.61742	116.3031966	0	0	218	3828546399278653	9级风过后的北京早晨天气真棒北京的好天气还真是风吹出来的 http://t.cn/z8ASTvL

3　微博特征地理规律性构建指标和过程

3.1　微博地理规律性的构建指标

为了能够对微博事件进行检测，必须在已获得的微博数据的基础上评估微博的地理规律性指标。本文主要针对微博数量、微博被转发数量、微博被评论数量、用户活跃强度、用户移动强度 5 个指标对微博的地理规律性进行构建，如图 3 所示:

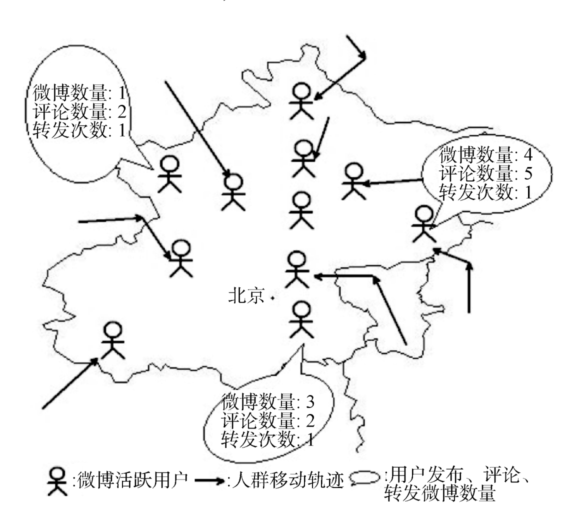


图 3　微博地理规律性指标

指标制定的依据如下:

(1) 微博发布数量：微博用户在公开表达动机的驱使下，会通过微博针对某事件公开发表言论和宣泄自己的情绪。此外，相当一部分用户受到社会提升动机的影响，为了获得更多关注和认可，吸引更多的粉丝，它们会积极更新微博的内容。因此，在特定的时间段里，该区域内的微博发布数量很有可能会偏离正常数量[15]。

(2) 用户活跃强度：指在一定的时间范围内发布至少一条微博的人数的总和。突发性的事件发生后，微博用户在公开表达和社会提升动机的驱使下，会纷纷创作发布关于该事件的微博内容。用户活跃强度很有可能会偏离正常水平。

(3) 微博被转发数量：由于国内外各种因素的影响，决定了社会上存在着诸多矛盾，这些矛盾往往会导致人为性突发事件(如恐怖主义、抢劫等)的发生。除此之外，还有一些自然性的突发事件(如地震、洪水等)。突发性事件是社会舆论关注的焦点，微博用户会对反映突发性事件的微博进行转发，因此转发数量很有可能偏离正常水平。青海玉树的大地震中，有微博用户通过微博发布救援信息:“玉树地震灾区靠西 100 公里有个叫隆宝镇的地方受灾严重，目前尚无救援队伍抵达”。这条微博经过微博用户的大量转发，最终引起了相关政府部门的重视，有效弥补了主流媒体的信息盲点[16]。

(4) 微博被评论数量：对于本地区突发性事件的微博，当地微博用户会对其进行持续的关注，同时往往会对此表达自己的观点、建议和情感等。此外，用户在转发微博时，也会对微博进行评论，这就会导致该条微博将获得大量的用户评论。

(5) 用户移动强度：人群的移动往往和一定的突发性事件有着密切的关系。按照用户移动类型可以分为三种：移入、移出和本地移动。移入是指人群从其他区域涌入本区域，移出是指人群从本区域移到其他区域，本地移动是指在本区域内移动，通常本地移动往往反映的是人们日常规律性的移动轨迹。当突发性事件发生时，用户的移动强度会处于一个非正常的区间内。比如玉树大地震时，有大量的人群为了逃避灾难，纷纷移出到其他区域；武汉大学赏樱期间，每天有 10-20 万人涌入校园，这说明用户移动强度对于事件检测有重要的意义[17]。

3.2　微博地理规律性的构建

(1) 微博特征的时间分布

根据微博数据中心发布的《2014 年微博用户发展报告》[18]的数据显示，微博用户每日微博发布、转发、评论行为在时间上存在极大的相似性，如图 4 所示。

三种微博用户随时间变化的趋势大致如下：从 0 点–5 点，用户基本都处于睡眠状态，三种用户行为的数量都呈下降趋势，处于低水平状态。从 5 点–11 点，微博用户逐渐活跃起来，三种用户行为呈逐步上升趋

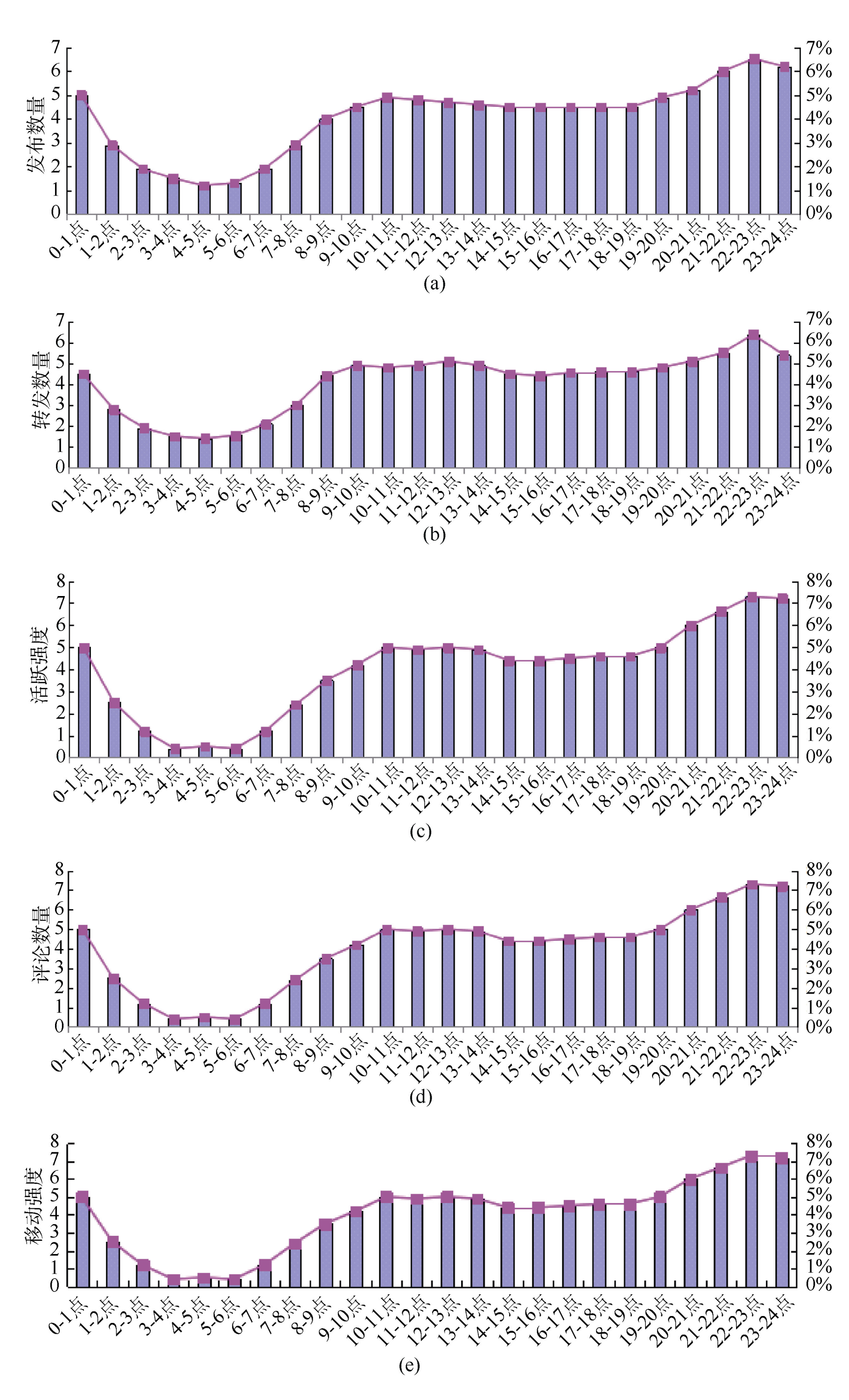


图 4　微博发布、转发、评论、活跃强度、移动强度的时间分布

势。从 11 点–18 点，三种用户行为略有波动，总体上处于高水平的稳定状态。从 18 点–24 点，用户往往结束了一天的工作，处于放松休息状态，有足够的时间和精力来发布、转发、评论微博，三种用户行为又呈现出了上升趋势，在 22 点–23 点，三种用户行为均达到峰值。在这一时段，用户的微博发布数量占全天微博发布总量的 6.53%，转发微博数量占全天微博转发总量的 6.37%，评论微博数量占全天微博评论总量的 7.61%。因此，本文按照以上 4 个时段，对一天的微博数据进行划分，进而发现各个时段的微博特征。此外，微博用户活跃强度和人群移动强度也与用户的作息时间息息相关，具有类似的时间分布特性。

(2) 微博特征的空间分布

由于自然条件、经济发展、文化传统的不同，不同地域的微博用户行为存在显著性的差异，如图 5 所示。北上广以及江浙地区这些经济实力雄厚，人口密度大的省份，微博用户分布较为密集且活跃，产生了大部分的微博数据。10 省的微博月活跃人数之和达到了 45.6%。

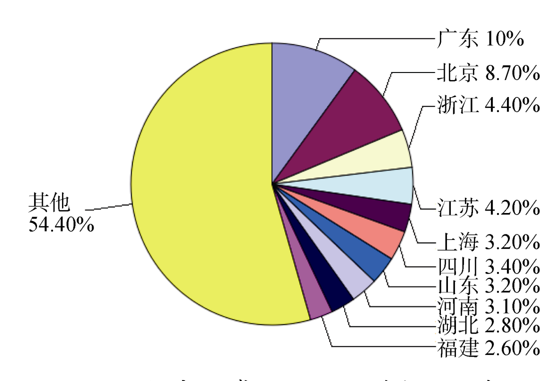


图 5　新浪微博月活跃人群省份分布

为了降低地区差异性的影响，本文采用 K-means 聚类方法[19]，使得同类中的微博地区差异性小，不同类中的微博地区差异性较大，达到高内聚、低耦合的效果。K-means 算法存在两个缺陷：聚类中心的个数 K 需要用户事先给定，但应该把数据对象分成多少个类别才最合适是无法事先确定的；K-means 需要人为地或者算法随机地确定初始聚类中心，不同的初始聚类中心可能导致完全不同的聚类效果。为了减轻两大缺陷的影响，本文提出如下两个对策：通过多次对微博数据聚类效果进行评价，得出一个实验性的 K 值，用该值作为聚类个数；可以借鉴 K-means++算法的思想在某种程度上解决随机性的问题。结合微博用户的区域性的特点，城市政府部门所在点经济发达、人口密度大，微博用户活跃度高，成为聚类中心的可能性最大，因此可以将其作为初始聚类中心之一，按照 K-means++的初始聚类中心之间的相互距离要尽可能远的原则，确定其他 K–1 个聚类中心。通过使用改进的 K-means 算法对表 2 中的数据按照经度和纬度对收集到的微博数据进行聚类分析，最终得到 K 个聚类中心点。

(3) 微博的地理规律性构建过程

通过时间和空间维度的处理，削弱了时空差异性对微博数据特征的影响，但是样本数据的收集期间，很可能也发生了一些突发事件，消除这些突发事件的影响，对于总结微博的地理规律性特征是至关重要的。

使用箱线图排除这些异常点的影响。箱线图反映了数据资料的最大值、上四分位数、中位数、下四分位数、最小值 5 个统计量，此外箱线图还反映出资料中的异常值。这里借用常用的异常值判断标准，将数据资料中超过上四分位数 1.5 倍四分位距(上四分位数+1.5×四分位距)或者低于下四分位数 1.5 倍四分位距(下四分位数–1.5×四分位距)的数据判定为异常值。通过 5 个指标的箱线图，去除异常值，将最大值和最小值之间的微博数据视为正常值，用来计算 5 个指标正常的数量水平，即微博的地理规律性特征。与图 3 相比，图 6 中微博的发布、转发、评论及微博用户活跃强度和用户移动强度都偏离了正常水平，因此要把这样的异常值去掉。图 6 中显示了时间段 11 点–18 点的微博异常数据的处理结果。

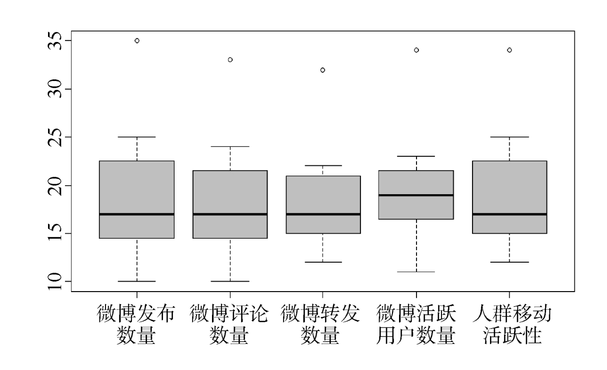


图 6　微博数据去除异常值

针对预处理过后的微博数据，就可以抽取五维指标的地理规律性特征：微博的发布数量通过分别累加特定空间的各个时间段的微博数量得到；微博转发、

评论数量通过累加特定空间的各个时间段的所有微博的转发数和评论数获得; 用户活跃度通过累加特定空间的各个时间段发布微博的用户数量计算, 某用户在该时间段内不管发布多少条微博, 都将活跃度记为1; 通过用户前后两次发微博的经度和纬度值计算得到用户移动距离, 将用户的移动距离的累加之和作为用户移动强度。

以收集到的连续一周内的数据做为训练样本, 对数据样本进行 K-means 聚类, 将所有的微博数据划分到对应的类内, 得到 K 个聚类中心, 将每个聚类内的微博数据按照 0 点–5 点、5 点–11 点、11 点-18 点和 18 点–24 点这 4 个时段进行划分, 得出各个时段内的用户微博发布、转发、评论数量、微博活跃用户数量和人群移动活跃性这 5 个指标的微博地理规律性特征。

4 微博事件检测框架

4.1 微博数据边界划分

对于日常的微博数据, 需要将其划分到对应的类中, 从而才能与对应的规律性进行比较。对空间边界进行划分, 常用的方法是维诺图[20](Voronoi Diagram)。基本思想是根据聚类得到的 K 个聚类中心, 采用维诺图对地理空间进行硬性的划分出多边形边界, 如果微博数据经纬度落在多边形范围内, 就将该微博数据判定到这个聚类中。该算法要想判定微博数据的归类情况, 需要比较大量的划分边界, 而且该方法可能把距离某个聚类中心很近的点, 划分到另外一个多边形中, 对微博事件的检测效果产生影响。因此, 本文采用经典的 KNN[21]最近邻分类算法, 通过投票的方式将待分类数据划分到相应的类中, 从而实现对微博数据边界的逻辑上的划分。

4.2 微博事件检测

微博事件检测的具体流程如下:

(1) 将前期收集到的一周内的带有经度和纬度的微博数据作为 K-means 聚类算法的输入, 最终得到 K 个聚类, 即划分除了 K 个空间。

(2) 将每个聚类中的数据按照时间段分为 4 部分, 去除训练样本集中的异常时间段, 目的是为了排除已发微博事件对微博特征的地理规律性造成干扰。

(3) 针对训练样本数据集, 抽取微博发布、转发、评论以及微博用户活跃度和用户移动强度 5 个指标上的规律性特征, 作为比较的标准, 将其保存在微博地理规律性数据库中。

(4) 随着时间的推移, 原有的微博规律性特征可能会过时, 会增加对微博事件误判的概率, 因此本文设定了一定的过期时间, 到期以后对整个的微博规律性特征进行重构。

(5) 通过微博数据采集程序, 从新浪微博每天收集待测数据, 存放到本地数据库中。

(6) 从本地数据库读取数据, 采用 KNN 最近邻分类算法对待测数据进行分类, 抽取出各个聚类中待测数据的微博发布、转发、评论以及微博用户活跃度和用户移动强度 5 个指标上的特征。

(7) 将步骤(3)中的微博地理规律性特征与步骤(6)中的日常微博地理特征进行比较, 审查各个指标是否偏离了正常水平, 从而判断是否发生了重大事件, 如果发生了重大事件, 则发出事件预警, 否则不做任何处理。

4.3 微博事件分类

根据事件的发生过程、机理和性质, 将微博事件分为 5 类: 自然灾害类(如地震、洪水等)、公共卫生事件类(如食品安全、动物疫情等)、社会安全事件类(恐怖袭击、群体性事件)、事故灾害类(如环境污染、煤矿坍塌等)、娱乐休闲类(如明星丑闻、电视影评等)。通过微博事件检测框架发现指标不正常的聚类, 人工阅读聚类中的微博内容, 判定这个聚类内的事件类型并予以标记。当同一聚类发生同类型事件时, 微博用户对事件的反映在很大程度上是类似的。因此, 当获得一定数量的带有标记的聚类数据(见表 3)后, 可以采用决策树[22](Decision Tree)的分类方法对后续的微博事件进行事件类型预测。

表 3 事件标记的微博数据

聚类 ID	发布数量	评论数量	转发数量	活跃人数	移动强度	事件类型
11	6891986	2414171	6790904	4571025	5805603	自然灾害
23	6175992	4852622	3375413	2667451	4079429	公共卫生
54	9202388	6691605	3127106	4605503	8500567	事故灾害
67	4442551	5477037	7880400	1586587	4663537	娱乐休闲
73	9416855	1358961	8927051	8459759	5438189	社会安全
…	…	…	…	…	…	…

4.4 微博事件摘要抽取

通过微博事件分类框架, 可以大致了解发生了什

么类型的事件。要想知道具体发生了什么事件，就需要对不正常聚类内的微博数据进行摘要提取操作。本文抽取热度高的 5 条微博返回，作为对应事件类型的摘要，事实证明该种方法是简单而有效的。微博的热度可以通过博文的评论数、转发数和用户的粉丝数计算。微博的评论数越多，说明存在大量微博用户针对微博内容进行激烈的讨论，表达自己的观点看法。如果用户对微博内容感兴趣，会对其进行转发，用户转发行为可以衡量微博热度蔓延的强度，而微博评论数、转发数与用户粉丝数之间存在密切联系，用户粉丝数的多少会影响用户评论和转发的数量。微博用户的粉丝数符合幂律分布，少部分用户拥有大量的粉丝，用户粉丝可以实时接受被关注者发布的信息，因此粉丝越多的用户发布的信息影响力更大。基于微博的转发数、评论数和用户的粉丝数，本文提出微博的热度计算公式如下：

$$Hot(W_i)=\alpha R_{wi}^{1/2}+\beta C_{wi}^{1/2}+\kappa \log(FL_{wi}+1) \quad (1)$$

其中，α、β 和 κ 是三个权重常数，α、β 取值为 2，κ 取值为 1，R_{wi} 是微博 W_i 一天内被转发的次数，C_{wi} 是微博 W_i 一天内被评论的次数，FL_{wi} 是微博用户粉丝数。根据微博热度计算公式，可以计算出每一条原始微博的热度值，按照热度值排序，返回热度最高的 5 条微博，作为微博事件的摘要。

5 实验分析及其改进

5.1 实验结果数据分析

本文将收集到的2015年4月15日的微博数据使用KNN 算法分到相应的聚类中，计算微博数据的发布数量、评论数、转发数以及用户活跃度和移动强度5项指标，与所在聚类中的微博数据的地理规律性特征进行比较，结果发现存在异常的聚类，在18点–24点时间段内，微博发布数量和微博用户活跃度都高于正常水平，如图7和图8所示。用户移动强度低于正常水平，用户微博评论和转发数量高于中位数。

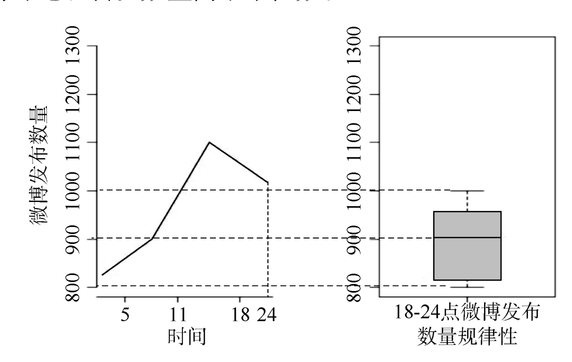


图 7 微博发布数量特征比较

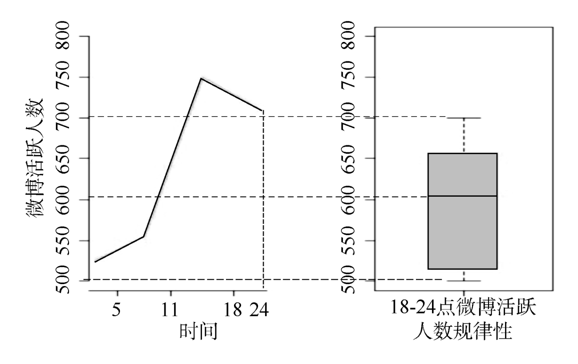


图 8 微博活跃用户数量比较

表 4 微博事件摘要

微博 TD	转发数	评论数	粉丝数	热度值	微博内容
3831952493218651	6032	2777	4993730	267.425	#北京沙尘暴#[沙尘暴入京了[衰]！]北京市气象台 15 日 17 时 40 分升级发布沙尘暴黄色预警信号，预计傍晚到夜间，本市将出现沙尘暴天气，能见度小于 1000 米，注意防范！好像电影《星际穿越》世界末日的即视感有木有！图自网友。
3832015281202961	3432	1334	34385739	197.751	[7 秒!看沙尘暴如何“吞没”CBD[衰]]今天，北京遭遇近 13 年来最强沙尘天气，北京商务中心区(CBD)8 分钟内被强沙尘笼罩，天空变黄变暗，能见度迅速降到 1 公里以下!把这 8 分钟缩成 7 秒，见识沙尘暴的厉害：秒拍视频 今天，你被沙子“侵袭”了吗?
3831970695789553	3344	1253	7756762	193.340	大家注意安全! [现场视频：北京沙尘暴肆虐 能见度低白昼瞬间变黑夜]
3831953021194620	2982	879	31435947	176.009	[漫天黄沙，此时的北方]正实时播报北方部分地区遭遇沙尘暴：http://t.cn/RA90Ch8
3831961162056497	1797	1033	40594335	156.671	[[话筒]沙尘“吞”北方 11 省区市！北京发大风沙尘双预警]今天北京、天津、河北、新疆、内蒙古、甘肃、宁夏、陕西、山西、辽宁、吉林等地有扬沙或浮尘，局地沙尘暴。目前北京已发大风黄色预警和沙尘蓝色预警，预计今晚京城有六七级风，阵风 9 级并伴沙尘。今天，你那儿吹沙了吗？

为了进一步了解到该聚类中的区域发生了什么事件，根据公式(1)计算数据库中每条微博的热度值，按照热度值高低进行排序，返回热度值最高的 5 条微博作为微博事件的摘要(见表 4)。通过阅读摘要，用户就能够及时发现微博事件，从而辅助事件相关部门以及个人提前作出决策，减小事件带来的负面影响。

通过表 4 中的微博发现，北京地区遭遇了恶劣的沙尘暴天气，因此北京微博用户都发表了大量的原创性微博，对沙尘暴天气进行跟踪报道，同时北京地区微博用户表达了对恶劣天气的不满情绪，因此微博发布数量和用户活跃度超过正常水平。

从图 9 中，可以看到在 18 点–24 点用户移动强度低于正常值。由于该时段北京地区正在遭遇沙尘暴天气，能见度极低，首要污染物从 PM2.5 变为 PM10。市环保监测中心数据显示，18 时开始，多个站点的 PM10 每小时浓度直线上升。北京监测网络 35 个站点 PM10 浓度均超过 1000 微克/立方米，达重度污染。在这种极其恶劣的天气情况下，北京地区的微博用户会避免外出活动，这就造成了用户移动强度低于正常水平。

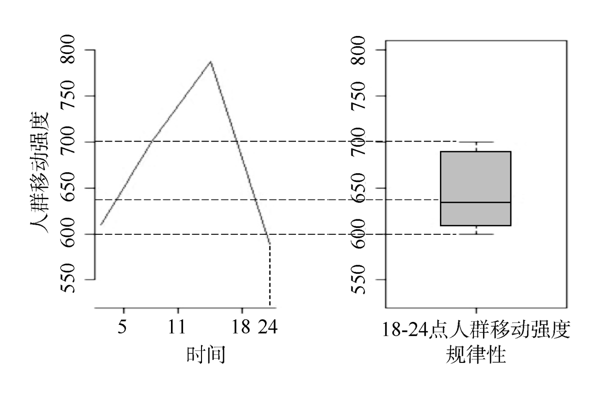


图 9　微博用户移动强度比较

从图 10 和图 11 可以看到，用户评论数和转发数均高于中位数，但是仍在正常值范围内。微博用户看到关于沙尘暴的微博，同时由于自己身临其境，正在经历沙尘暴，这就大大增加了用户对沙尘暴微博进行评论和转发的概率。微博用户在信息分享和公开表达动机的驱使，纷纷对沙尘暴天气的微博进行评论或转发评论，表达自己的感受、情绪和意见等，这就造成了用户评论数和转发数高于中位数的现象。

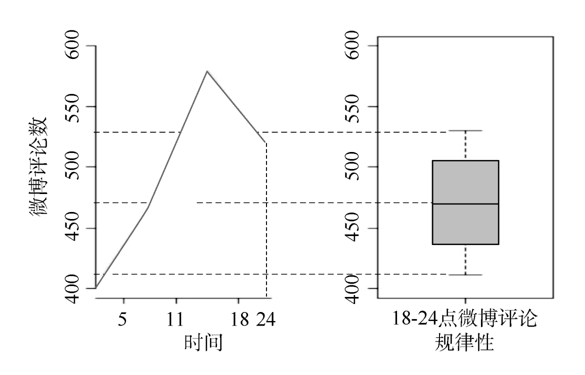


图 10　微博用户评论数量比较

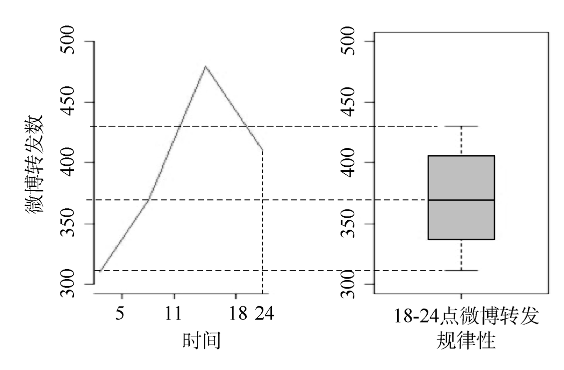


图 11　微博用户转发数量比较图

5.2　结果改进分析

为了更加清晰直接地显示北京沙尘暴事件的发展趋势，本文通过改进迟呈英等[23]提出的“话题指数”，对北京沙尘暴事件的生命周期进行分析。与网络新闻热点话题相比，微博事件具有周期性短的特点，因此本文对话题指数进行改进，将时间间隔设定为小时，相应提出了微博事件指数(Event Index)的概念。将微博指数定义为每小时内微博发布数增长量与第一小时内的微博增长量比值的权重函数，计算公式如下：

$$EI(E_i)=(P_{Ei}(t_{j+1})-P_{Ei}(t_j))\times P_{base}/P_{Ei}(t_1) \quad (2)$$

其中，E_i 表示某个微博事件，$P_{Ei}(t_j)$表示从初始时刻到 t_j 时刻的与事件 E_i 相关的微博累积量，$P_{Ei}(t_{j+1})-P_{Ei}(t_j)$表示 t_{j+1} 与 t_j 时刻之间事件相关微博发布的数量。$P_{Ei}(t_1)$表示初始的第一个小时内事件相关微博的发布数量。P_{base} 为微博事件出现第一小时的事件指数，给定 $P_{base}=1$。如果用横坐标表示时间，以小时为间隔，纵坐标表示微博事件指数，可以得到一条连续的曲线，称为微博事件发展趋势曲线。

虽然微博事件在某个单一聚类中被检测到，但是微博事件在现实世界中会对周围区域产生一定的影响，因此在分析微博事件发展趋势时，不能局限在单

一聚类中，而应该将范围扩展到整个受影响的区域。北京微博用户在2015年4月15日总共发布了以“沙尘暴”为主题的微博 21 966 条，将微博数据以小时为间隔进行切分，通过公式(2)计算每小时内的微博事件指数，如表5所示：

表 5　北京沙尘暴微博事件指数

时间	微博指数(4月15日)	微博指数(4月16日)	时间	微博指数(4月15日)	微博指数(4月16日)
0点-1点	1	114.75	12点-13点	16.25	48.75
1点-2点	1	32	13点-14点	7.5	42.75
2点-3点	0	14.75	14点-15点	14.5	38.75
3点-4点	0	7.25	15点-16点	9.25	38.75
4点-5点	0.5	8.25	16点-17点	33.75	26.75
5点-6点	0	8.25	17点-18点	498.25	23.25
6点-7点	2	37.75	18点-19点	2118.25	29.5
7点-8点	2.5	65.5	19点-20点	927.5	23
8点-9点	10.75	101.5	20点-21点	652.5	16.25
9点-10点	30	105	21点-22点	477.75	20.25
10点-11点	29.75	85	22点-23点	376.75	24
11点-12点	12.25	65	23点-24点	269.5	18.5

根据表5中的数据，采用R语言作图对沙尘暴事件发展趋势进行可视化分析。从图12中可以发现，4月15日0点–9点，由于沙尘暴尚未发生，对于沙尘暴事件只有极少量用户关注和讨论，此时处于事件的潜伏期[24]。9点–17点，沙尘暴事件进入萌动期，此阶段沙尘暴事件已经初现端倪，微博事件指数呈现出一定的波动状态。17点–18点30分，微博用户对沙尘暴事件的关注迅速提高，沙尘暴事件的影响范围借助微博平台急剧扩张，微博事件进入加速期，这一剧烈变化与沙尘暴的爆发时间是吻合的。18点30分–19点30分，沙尘暴的时间指数的提高速率放缓，在19点时微博事件指数达到最大值2 118.25，沙尘暴事件进入成熟期。微博的传播特点决定了其用户的注意力必然是有限且多变的，呈现出碎片化与表面化的趋势，所以很难在微博上对某一事件进行长期深入的关注[25]。从2015年4月15日19点30分–4月16日24点，随着沙尘暴的过境，对用户生活的影响力减小，用户对沙尘暴的关注力转移到其他的事件上，沙尘暴事件进入到了衰退期。4月16日6点–15点，事件指数值出现了轻微的波动趋势。通过阅读微博内容可以了解到，4月16日，北京重现蓝天白云，微博用户纷纷对比两天的天气状况，发表了一定数量的微博，造成了时间指数的起伏的状况。

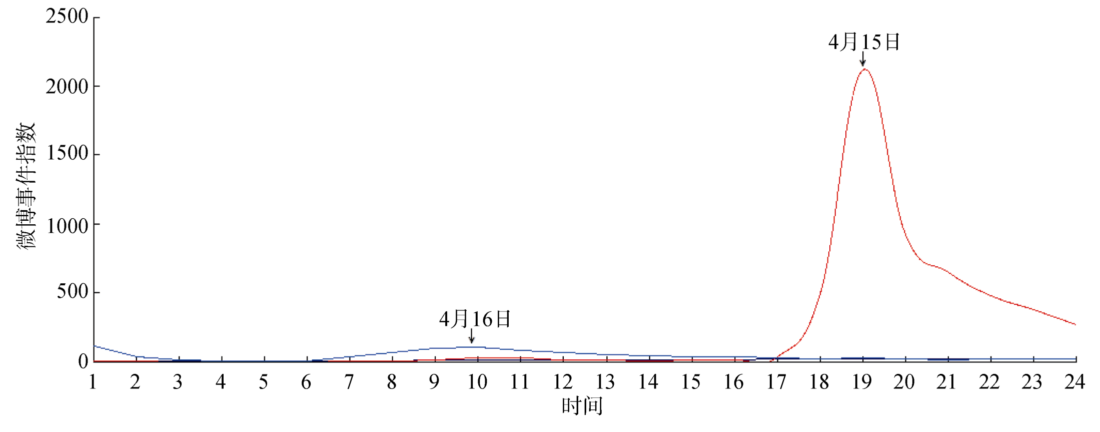


图 12　沙尘暴微博事件发展趋势

6　结　语

本文设计一种基于 IP 轮替的多用户的微博数据采集方案，实现微博数据的高效采集功能。针对采集到的微博数据，制定了微博发布数量、微博转发和评论数量、微博用户活跃度和移动强度 5 个指标，衡量微博的地理规律性特征。本文利用K-means聚类、KNN分类和决策树三种数据挖掘算法，提出一个详细的微博地理规律性抽取架构，设计并实现了微博事件检测功能。通过微博事件检测实验，验证了该微博事件检测方法的有效性。

本文提出的微博事件检测架构也存在一些需要改进的地方。由于实验条件的限制，在抽取微博地理规律性特征时，采用的样本数据偏少，一定程度上影响了事件检测框架的效果；可视化环节需要人工的干预，没有实现完全的自动化处理流程。如果要对全国范围内的突发事件进行检测，就要处理海量的微博数据，这将会影响到事件检测的效率。为了解决以上问题，未来计划采用基于分布式文件存储和计算的 Hadoop[26]平台进行事件检测系统的搭建。可以在 Hadoop 的 HDFS (Hadoop Distributed File System)文件系统的基础上，部署 HBase 数据库，进行微博大数据的存储，利用 Hadoop 的 MapReduce 进行微博事件检测算法的实现，从而实现全国范围内的并且高效率的微博事件检测。本文采用 K-means、KNN 和决策树三种算法对微博数据进行分析研究，Hadoop 生态圈里的开源项目 Mahout 已经将三种算法 MapReduce 化，可以很方便地使用。对于微博事件可视化模块，可以使用 R 语言和 Hadoop 的结合产物 RHadoop 完成，最终实现可视化模块的自动化处理。

参考文献：

[1] 胡吉明. 社会化网络服务的开放运行架构及服务拓展研究[J]. 情报科学, 2012, 30(9): 1396-1400. (Hu Jiming. Study on Open Operation Architecture and Service Expansion of Social Network Service [J]. Information Science, 2012, 30(9): 1396-1400.)

[2] 李彪. 微博意见领袖群体“肖像素描”——以 40 个微博事件中的意见领袖为例[J]. 新闻记者, 2012(9):19-25. (Li Biao. The “Portrait Sketch” of Microblogging Opinion Leaders Group——Take 40 Opinion Leaders from Microblogs as an Example [J]. Journalism Review, 2012(09): 19-25.)

[3] 杨亮, 林原, 林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报, 2012, 26(1):84-90. (Yang Liang, Lin Yuan, Lin Hongfei. Micro-Blog Hot Events Detection Based on Emotion Distribution [J]. Journal of Chinese Information Processing, 2012, 26(1): 84-90.)

[4] 王林, 时勘, 赵杨, 等. 基于突发事件的微博集群行为舆情感知实验[J]. 情报杂志, 2013, 32(5): 32-37. (Wang Lin, Shi Kan, Zhao Yang, et al. Experimental Studies on Public Opinion Perception of the Micro Blog’s Collective Behavior Based on the Emergencies [J]. Journal of Intelligence, 2013, 32(5): 32-37.)

[5] 杨娟娟, 杨兰蓉, 曾润喜, 等. 公共安全事件中政务微博网络舆情传播规律研究——基于“上海发布”的实证[J]. 情报杂志, 2013, 32(9):11-15. (Yang Juanjuan, Yang Lanrong, Zeng Runxi, et al. Research on Communication Mechanism of Internet Public Opinion of Government Affairs Microblog in Public Security Events: A Case Study of the “Shanghai Fabu” [J]. Journal of Intelligence, 2013, 32(9): 11-15.)

[6] 兰月新. 突发事件微博舆情扩散规律模型研究[J]. 情报科学, 2013, 31(3): 31-34. (Lan Yuexin. Research on Microblog Opinion Diffusion Model of Emergent Events [J]. Information Science, 2013, 31(3): 31-34.)

[7] 王勇, 肖诗斌, 郭跚秀, 等. 中文微博突发事件检测研究[J]. 现代图书情报技术, 2013(2): 57-62. (Wang Yong, Xiao Shibin, Guo Yixiu, et al. Research on Chinese Micro-blog Bursty Topics Detection [J]. New Technology of Library and Information Service, 2013(2): 57-62.)

[8] 陈国兰. 基于爆发词识别的微博突发事件监测方法研究[J]. 情报杂志, 2014, 33(9): 123-128. (Chen Guolan. Micro-blog Emergencies Detection Approach Based on Burst Words Distinguishing [J]. Journal of Intelligence, 2014, 33(9): 123-128.)

[9] 刘志明, 刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程, 2011, 29(6): 8-16. (Liu Zhiming, Liu Lu. Recognition and Analysis of Opinion Leaders in Microblog Public Opinions [J]. Systems Engineering, 2011, 29(6): 8-16.)

[10] 魏志惠, 何跃. 基于信息熵和未确知测度模型的微博意见领袖识别——以“甘肃庆阳校车突发事件”为例[J]. 情报科学, 2014, 32(10): 38-43. (Wei Zhihui, He Yue. Identify Microblogging Opinion Leaders Based on Information Entropy and Unascertained Measure Model——Taking “Emergencies of Qingyang School Bus” as an Example [J]. Information Science, 2014, 32(10): 38-43.)

[11] 田野. 基于微博平台的事件趋势分析及预测研究[D]. 武汉: 武汉大学, 2012. (Tian Ye. On Trends Analysis and Prediction Based on Micro-Blogging Platforms [D]. Wuhan: Wuhan University, 2012.)

[12] Yang Y, Carbonell J, Brown R. Multi-Strategy Learning for Topic Detection and Tracking [A]. // Topic Detection and Tracking [M]. Springer, 2002: 85-114.

[13] 冯永, 韩楠, 贾东风. 云计算环境下基于代表点增量层次密度聚类的微博事件检测及跟踪[J]. 计算机应用, 2013, 33(12): 3559-3562. (Feng Yong, Han Nan, Jia Dongfeng. Microblog Events Detection and Tracking with Incremental Hierarchical DBSCAN Based on Representative Posts Using

Cloud Framework [J]. Journal of Computer Applications, 2013, 33(12): 3559-3562.)

[14] 王连喜. 微博短文本预处理及学习研究综述[J]. 图书情报工作, 2013, 57(11):125-131. (Wang Lianxi. A Literature Review on Pre-processing and Learning of Microtext [J]. Library and Information Service, 2013, 57(11): 125-131.)

[15] Fu C, Samet H, Sankaranarayanan J. WeiboStand: Capturing Chinese Breaking News Using Weibo "Tweets" [C]. In: Proceedings of the 7th ACM SIGSPATIAL Workshop on Location-Based Social Networks. 2014.

[16] 王锋. 灾难性事件中的"微"力量——青海玉树地震中微博应用探析[J]. 新闻世界, 2010(S2): 149-150. (Wang Feng. "Micro" Forces of the Catastrophic Event——Qinghai Yushu Weibo Application Analysis in the Earthquake [J]. News World, 2010(S2): 149-150.)

[17] Zhang P. Social Inclusion or Exclusion? When Weibo (Microblogging) Meets the "New Generation" of Rural Migrant Workers [J]. Library Trends, 2013, 62(1):63-80.

[18] 微博数据中心. 2014 年微博用户发展报告[R/OL]. [2015-02-06]. http://data.weibo.com/report/reportDetail?id=215. (Weibo Data Center. The 2014 Report of Weibo Users Development [R/OL]. [2015-02-06]. http://data.weibo.com/report/reportDetail?id=215.)

[19] 吴夙慧, 成颖, 郑彦宁, 等. K-means 算法研究综述[J]. 现代图书情报技术, 2011(5): 28-35. (Wu Suhui, Cheng Ying, Zheng Yanning, et al. Survey on K-means Algorithm [J]. New Technology of Library and Information Service, 2011(5): 28-35.)

[20] 亓峰, 刘昆, 张超, 等. 圆和维诺图相交模拟基站覆盖算法[J]. 北京邮电大学学报, 2014, 37(S1): 108-114. (Qi Feng, Liu Kun, Zhang Chao, et al. A Novel Base Station Coverage Simulation Based on Intersection of Circle and Voronoi[J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(S1): 108-114.)

[21] 江涛, 陈小莉, 张玉芳, 等. 基于聚类算法的 KNN 文本分类算法研究[J]. 计算机工程与应用, 2009, 45(7): 153-158. (Jiang Tao, Chen Xiaoli, Zhang Yufang, et al. Improved KNN Using Clustering Algorithm [J]. Computer Engineering and Applications, 2009, 45(7): 153-158.)

[22] 陆安生, 陈永强, 屠浩文. 决策树 C5 算法的分析与应用[J]. 电脑知识与技术, 2005(3): 17-20. (Lu Ansheng, Chen Yongqiang, Tu Haowen. The Analysis and Application of Decision Tree Algorithm of C5 [J]. Computer Knowledge and Technology, 2005(3): 17-20.)

[23] 迟呈英, 李红. 基于改进 TF*PDF 算法的网络新闻热点话题检测和跟踪[J]. 计算机应用与软件, 2013, 30(12): 311-314. (Chi Chengying, Li Hong. Network News Hot Topics Detection and Tracking Based on Modified TF*PDF Algorithm [J]. Computer Applications and Software, 2013, 30(12): 311-314.)

[24] 谢科范, 赵湜, 陈刚, 等. 网络舆情突发事件的生命周期原理及集群决策研究[J]. 武汉理工大学学报: 社会科学版, 2010, 23(4): 482-486. (Xie Kefan, Zhao Shi, Chen Gang, et al. Research on Lifecycle Principle and Group Decision-making of Network Public Sentiment Emergency [J]. Journal of Wuhan University of Technology: Social Sciences Edition, 2010, 23(4): 482-486.)

[25] Narayanam R, Narahari Y. A Shapley Value-based Approach to Discover Influential Nodes in Social Networks [J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(1): 130-147.

[26] 陈吉荣, 乐嘉锦. 基于 Hadoop 生态系统的大数据解决方案综述[J]. 计算机工程与科学, 2013, 35(10): 25-35. (Chen Jirong, Le Jiajin. Reviewing the Big Data Solution Based on Hadoop Ecosystem [J]. Computer Engineering & Science, 2013, 35(10): 25-35.)

作者贡献声明:

李进华: 提出和论证研究命题, 部分内容撰写及最终版本修订;
安仲杰: 数据收集与分析, 提出实验验证方案, 修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-2]由作者自存储, 可通过电子邮件向作者索取, E-mail: anzhongjie01@163.com; 支撑数据[3-11]见期刊网络版 http://www.infotech.ac.cn。

[1] 李进华, 安仲杰. rawdata.txt. 微博原始数据.

[2] 李进华, 安仲杰. classdata.txt. 微博聚类结果数据.

[3] 李进华, 安仲杰. GetNearbyGeo.java. 根据经纬度获取微博数据代码. http://pan.baidu.com/s/1mgZ4vUS.

[4] 李进华, 安仲杰. Map.java. 计算用户移动强度代码. http://pan.baidu.com/s/1jHzqENG.

[5] 李进华, 安仲杰. box.txt. R 语言绘制对比图代码. http://pan.baidu.com/s/1pK46wSr.

[6] 李进华, 安仲杰. release.csv. 微博发布数量的时间分布数据.

[7] 李进华, 安仲杰. comment.csv. 微博评论数量的时间分布数据.

[8] 李进华, 安仲杰. forwarding.csv. 微博转发数量的时间分布数据.

[9] 李进华, 安仲杰. active.csv. 用户活跃强度的时间分布数据.
[10] 李进华, 安仲杰. move.csv. 用户移动强度的时间分布数据.
[11] 李进华, 安仲杰. index.csv. 沙尘暴微博事件指数.

收稿日期: 2015-09-24
收修改稿日期: 2015-12-04

Analyzing Geographical Coordinates Data for Micro-blog Trending Events

Li Jinhua　An Zhongjie
(School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: [**Objective**] This study aims to retrieve the trending events from the micro-blog platform with the help of data mining algorithms. [**Methods**] First, we collected micro-blog message with geographic coordinates from the most popular platform (the Sina Weibo) using its API service. Then, we used the K-means, KNN and decision trees algorithms to construct the geographical patterns of those collected posts. The number of published posts, re-tweets, and comments, as well as user activity and movement strength were also examined. Third, we compared these geographical patterns with the daily regional micro-blog data to identify breaking news in that area. [**Results**] We analyzed data collected on April 15 and April 16 of 2015 with the help of the proposed model, and found a trending event of "Beijing Sandstorm". [**Limitations**] The sample size was small, which might influence the results. [**Conclusions**] Geographic coordinates could help us detect trending events on the Sina Weibo, and this new method will also support the government's crisis management strategy and decision-making process.

Keywords: Micro-blog　Event detection　Visualization analysis　Geographical coordinates analysis